

Pattern recognition at different scales: a statistical perspective

Matteo Colangeli¹, Francesco Rugiano¹, Eros Pasero¹

¹*Dipartimento di Elettronica e Telecomunicazioni, Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Torino, Italy*

Abstract

In this paper we borrow concepts from Information Theory and Statistical Mechanics to perform a pattern recognition procedure on a set of x-ray hazelnut images. We identify two relevant statistical scales, whose ratio affects the performance of a machine learning algorithm based on statistical observables, and discuss the dependence of such scales on the image resolution. Finally, by averaging the performance of a Support Vector Machines algorithm over a set of training samples, we numerically verify the predicted onset of an “optimal” scale of resolution, at which the pattern recognition is favoured.

1. Introduction

The theory of artificial neural networks (ANN) represents an open research field setting the stage for the implementation of a statistical mechanical approach in novel interdisciplinary problems, such as the modeling of the collective behavior of the human brain neurons. An important field of application of ANN is represented by the pattern recognition analysis [1, 2], which has received an increasing interest in the literature, witnessed by the extensive application of ANN to tackle complex real-word problems, e.g. in medical diagnosis [3, 4, 5] and in biological sequences analysis [6, 7, 8, 9]. Recent works, in this field, paved also the way to the systematic use of technical tools borrowed from Information Theory and Statistical Mechanics [10, 11, 12].

In this paper, in particular, we adopt information theoretic methods [13, 14]

Email address: matteo.colangeli@polito.it, francesco.rugiano@polito.it, eros.pasero@polito.it (Matteo Colangeli¹, Francesco Rugiano¹, Eros Pasero¹)

to classify a sequence of hazelnuts images, and show how our approach allows for improving the performance of pattern recognition procedures performed via ANN algorithms. From a preliminary statistical analysis on the image histograms, we identify some relevant observables to be used in the implementation of a machine learning algorithm. A special focus of our approach is on the role of *fluctuations* of the histograms around the corresponding *mean* distribution. In particular, by making use of various notions of “distance” between histograms, we introduce two statistical scales, whose magnitude affects the performance of a machine learning algorithm in disentangling and extracting the distinctive features of the hazelnuts.

The paper is organized as follows.

In Sec. 2 we introduce the two aforementioned statistical scales and discuss their dependence on a quantity referred to as the “image resolution”. We comment on the need of a large separation between two such scales to obtain an efficient pattern recognition: the lack of a wide separation between them is due to large histograms fluctuations which blur the distinctive features of the hazelnuts, thus hindering a proper classification of the data.

In Sec. 3 we test, then, the prediction of our statistical analysis by employing a machine learning algorithm, known as Support Vector Machines (SVM) [15, 16]. The numerical results we obtained not only confirm the relevance of the aforementioned scale separation, but also show that the predicted onset of an optimal scale of description can be recovered through the use of a SVM algorithm, provided that its performance is *averaged* over a sufficiently large set of training samples.

Conclusions are finally drawn in Sec. 4.

The main results of this work can be summarized as follows:

- We introduce two typical statistical scales, whose magnitude critically affects the performance of a pattern recognition algorithm based on statistical variables;
- We describe the dependence of such scales on the scale of resolution, thus unveiling the onset of an optimal resolution at which the pattern recognition is favoured;
- We numerically recover the results of the statistical analysis by using a SVM algorithm, and also shed light on the role of *averaging* the performance of a SVM over sufficiently many training samples.

2. The original set of hazelnut images: a statistical approach

In this work we consider the problem of pattern recognition applied to a sequence of hazelnut images, to be categorized into three different sets: “good” (G), “damaged” (D) and “infected” (I). In the sequel, we will use the shorthand notation $\mathcal{S} = \{G, D, I\}$, and, for any $A \in \mathcal{S}$, we will also denote $N_A = \text{card}(A)$. Our database consists of a set of 800 x-ray scanned images, cf. Fig. 1, with $N_G = 750$, $N_D = 25$ and $N_I = 25$. The analysis outlined below is meant to provide a guiding strategy to assess, and possibly enhance, the performance of pattern recognition methods based on ANN algorithms. The prominent distinctive features of the three sets G , D and I are not detectable from a solely visual inspection of the x-ray images. Hence, in order to extract some valuable information, we relied on the computation of the histograms of the hazelnut images, shown in Fig. 2.

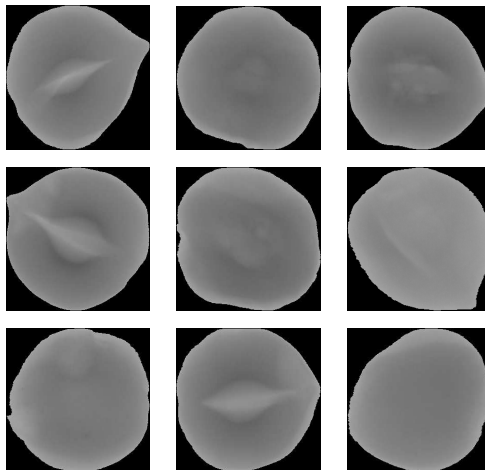


Figure 1: X-ray scanned images of good hazelnuts (top row), damaged hazelnuts (middle row) and infected hazelnuts (bottom row).

Therefore, for any $A \in \mathcal{S}$, we computed the number of pixels, in the image pertaining to the i -th hazelnut belonging to the set A (with $i = 1, \dots, N_A$), characterized by the shade of gray j (conventionally running from the value 0 - black - to 255 - white). After normalizing wrt the total number of pixels forming the same image, we thus obtained the so-called image histogram $p_i^{(A)}(j)$. We could also compute, then, the mean histogram pertaining to A , denoted by $\bar{p}_i^{(A)}(j)$, which was obtained by averaging over the N_A histograms $p_i^{(A)}(j)$.

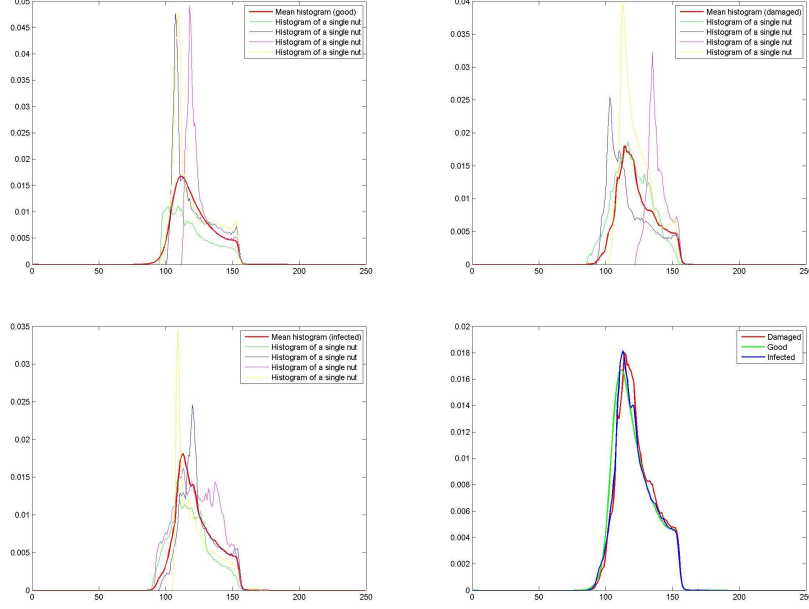


Figure 2: Image histograms of good hazelnuts (top left), damaged hazelnuts (top right), infected hazelnuts (bottom left). The horizontal axis displays the shades of gray, conventionally running from 0 to 255. For each of the three sets, the figures display the (normalized) histograms of single hazelnuts as well as the (normalized) mean histogram. On the bottom right corner, the mean histograms of the three different sets are compared.

A quantitative characterization of the images can be afforded by introducing various notions of “distance” between different histograms [17]: we considered, in particular, the norm in L^1 , in L^2 (euclidean), in L^∞ , the Squared χ^2 distance and the Jeffrey’s divergence [18]. It is worth briefly recalling some basic aspects concerning the latter two notions of distance, borrowed from probability theory. The Squared χ^2 distance corresponds to the symmetrized version of the Pearson’s χ^2 test [19], which, given a histogram $p(j)$ and a reference histogram $q(j)$, defines their relative distance as:

$$d_{\chi^2} = \sum_j \frac{(p(j) - q(j))^2}{q(j)} . \quad (1)$$

Thus, the quantity d_{χ^2} in (1) resembles the standard euclidean distance between the two histograms, except that it introduces a weight corresponding to the inverse of the reference histogram.

On the other hand, the Jeffreys’ divergence [20] belongs to the Shannon

entropy family [21], and corresponds to the symmetrized version of the Kullback-Leibler (K-L) divergence (or *relative entropy*) [22], defined as:

$$d_{K-L}(p||q) = \sum_j \left(p(j) \log \left(\frac{p(j)}{q(j)} \right) \right) = H(p, q) - H(p) \quad , \quad (2)$$

where $H(p, q)$ is the cross entropy of p and q , and $H(p)$ is the entropy of p [23, 24]. More in general, the K-L divergence (2), is a member of the family of the so-called f -divergencies [25, 26] and stems as a limiting case of the more general Rényi's (or α -) divergence [27]. It is worth recalling its definition: given any two continuous distributions p and q , over a space Ω , with p absolutely continuous wrt q , the f -divergence of p from q is

$$d_f(p||q) = \int_{\Omega} f \left(\frac{dp}{dq} \right) dq \quad , \quad (3)$$

where f is a convex function such that $f(1) = 0$.

Then, for any $A \in \mathcal{S}$, we considered the distance (or *fluctuation*), defined according to the various notions introduced above, between the histogram $p_i^{(A)}(j)$ and the corresponding mean $\bar{p}_i^{(A)}(j)$. Next, by averaging over the set A , one obtains a characteristic “statistical scale” (still depending on the chosen notion of distance) characterizing the fluctuations within each set A . To clarify the meaning of the entries in Tab. 1, let us illustrate, for instance, the procedure to calculate the quantity $\langle d \rangle_2^{(A)}$. To this aim, we introduce the euclidean distance between the histograms $p_i^{(A)}(j)$ and $\bar{p}_i^{(A)}(j)$:

$$d_{2,i}^{(A)} = \sqrt{\sum_{j=1}^{N_g} |p_i^{(A)}(j) - \bar{p}_i^{(A)}(j)|^2} \quad (4)$$

	$\langle d \rangle_1^{(A)}$	$\langle d \rangle_2^{(A)}$	$\langle d \rangle_{\infty}^{(A)}$	$\langle d \rangle_{\chi^2}^{(A)}$	$\langle d \rangle_J^{(A)}$
$A = G$	0.2079	0.0372	0.0139	0.0495	0.0369
$A = D$	0.2485	0.0488	0.0162	0.0776	0.0477
$A = I$	0.2097	0.0379	0.0145	0.0435	0.0401

Table 1: Typical fluctuation of the histograms of the hazelnuts from the corresponding mean histogram, within each of the sets G , D , and I . The quantities $\langle d \rangle^{(A)}$ are evaluated by using different notions of distances: norm in L^1 , in L^2 (euclidean), in L^∞ , Squared χ^2 distance and Jeffreys divergence.

	$\Delta_1^{(A,B)}$	$\Delta_2^{(A,B)}$	$\Delta_\infty^{(A,B)}$	$\Delta_{\chi^2}^{(A,B)}$	$\Delta_J^{(A,B)}$
$A = G, B = D$	0.0923	0.0162	0.0036	0.0089	0.0200
$A = D, B = I$	0.0533	0.0090	0.0028	0.0021	0.0030
$A = G, B = I$	0.0526	0.0115	0.0051	0.0044	0.0124

Table 2: Average distances between pairs of mean histograms referring to two different sets A and B , evaluated, as in Tab 1, using different notions of distance: norm in L^1 , in L^2 (euclidean), in L^∞ , Squared χ^2 distance and Jeffreys divergence.

From the knowledge of $d_{2,i}^{(A)}$ in (4), the quantity $\langle d \rangle_2^{(A)}$, shown in Tab. 1, is then computed by averaging over A :

$$\langle d \rangle_2^{(A)} = \frac{1}{N_A} \sum_{i=1}^{N_A} d_{2,i}^{(A)} \quad (5)$$

It is worth noticing, from Tab. 1, that, no matter of what notion of distance is adopted, the magnitude of the fluctuations is not significantly affected by N_A . The scale $\langle d \rangle^{(A)}$, which, for any $A \in \mathcal{S}$, is of the order $\langle d \rangle^{(A)} \simeq 10^{-2}$, can be thus regarded as an intrinsic statistical scale pertaining to the set A . It is worth comparing such scale with another statistical scale, denoted by $\Delta^{(A,B)}$, whose values are listed in Tab. 2. The quantity $\Delta^{(A,B)}$ is defined as the distance, computed by using the various notions of distance introduced above, between the pair of mean histograms relative to the sets A and B , with $(A, B) \in \mathcal{S}$ and $A \neq B$. The symmetric form of the distances introduced above entails, in particular, that $\Delta^{(A,B)} = \Delta^{(B,A)}$. A better interpretation of the meaning of the scales $\langle d \rangle^{(A)}$ and $\Delta^{(A,B)}$ can be achieved by noticing that a large value of $\langle d \rangle^{(A)}$ mirrors the presence of a considerable amount of noise on top of the mean histogram $\bar{p}_i^{(A)}(j)$, which thus blurs the distinctive features of the set A . On the contrary, a larger value of $\Delta^{(A,B)}$ reflects a more significant separation between the mean histograms of the two sets A and B , which instead favours the pattern recognition. In the sequel of this Section we will focus, therefore, on the ratio of two such scales. From an inspection of Tabs. 1 and 2, we first observe that $\Delta^{(A,B)} \sim \langle d \rangle^{(A)}$. That is, the two scales are comparable: the fluctuations, within each set, are comparable with the typical distances between different sets. This entails, hence, that the histograms shown in Fig. 2 can not be regarded as a useful source of information to perform a pattern recognition. A different route can be pursued by just focusing on a selected portion of the original images. This approach is motivated by the assumption that the distinctive features of each of the three sets are mostly contained in the “nuclei” of the hazelnuts. We

calculated, therefore, the histograms corresponding to the cropped portions of the original images, delimited by the tick red rectangles shown in Fig. 3.

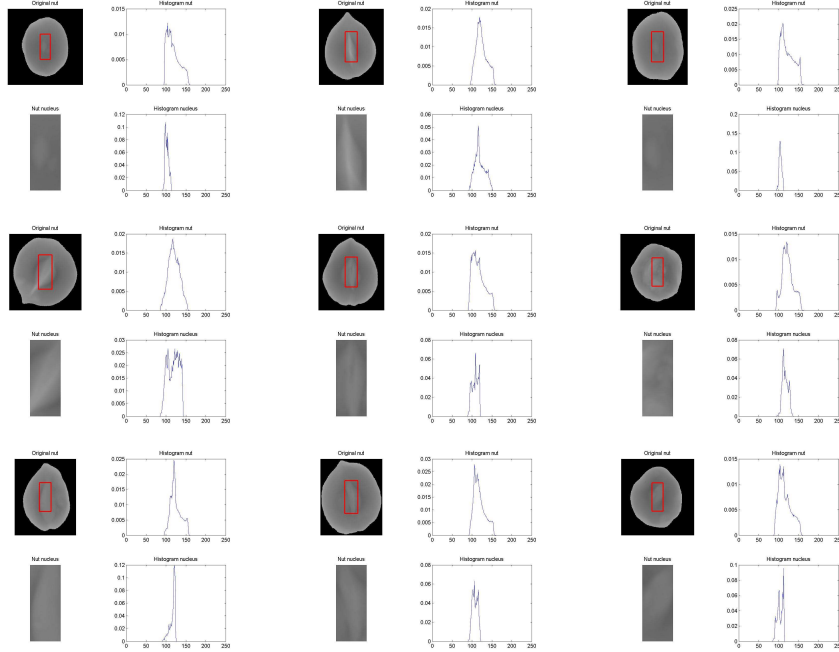


Figure 3: Image histograms of good hazelnuts (top row), damaged hazelnuts (middle row) and infected hazelnuts (bottom row). Each image shows the histogram of the entire hazelnut (top histogram) and the histogram referring to the fraction of the image delimited by the thick red rectangles, characterized by $\epsilon = 80$ and $\rho = 2.5$.

The red rectangles in Fig. 3 are identified by the pair of parameters $\{\epsilon, \rho\}$, where ϵ , related to the image resolution, is defined as the number of pixels comprised along the horizontal length of the rectangles, while ρ is the ratio of the number of pixels along the vertical length to the corresponding number of pixels along the horizontal one. In our simulations, the values of the parameters $\{\epsilon, \rho\}$ were kept constant when calculating the histograms relative to different hazelnut nuclei. Figure 3 refers, for instance, to the case corresponding to $\epsilon = 80$ and $\rho = 2.5$. In Figs. 4, 5 and 6, shown is the result of the image processing of the hazelnut nuclei, performed through a noise removal filter (adaptive Wiener filtering) and various edge-detector algorithms.

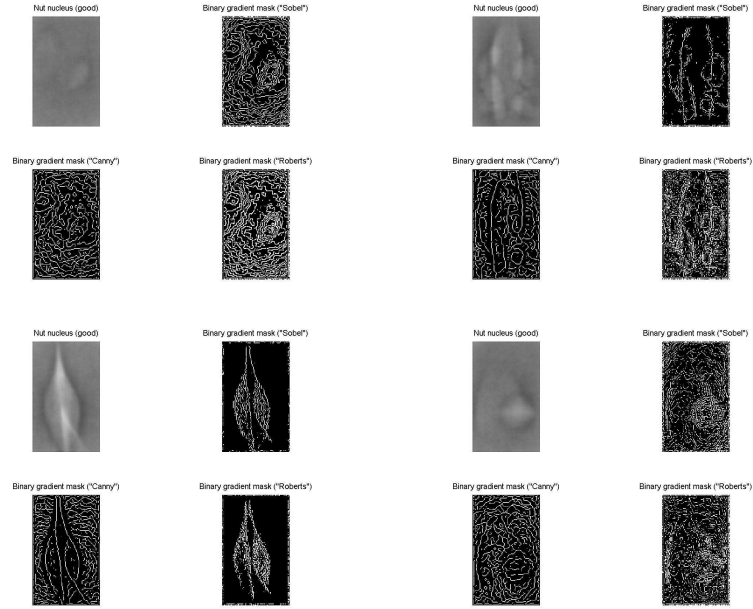


Figure 4: Image processing of the hazelnut nuclei belonging to the set G , for $\epsilon = 100$ and $\rho = 1.5$, by means of edge-detection algorithms, respectively: Sobel's algorithm (top right figure) , Canny's algorithm (bottom left figure) and Roberts' algorithm (bottom right figure).

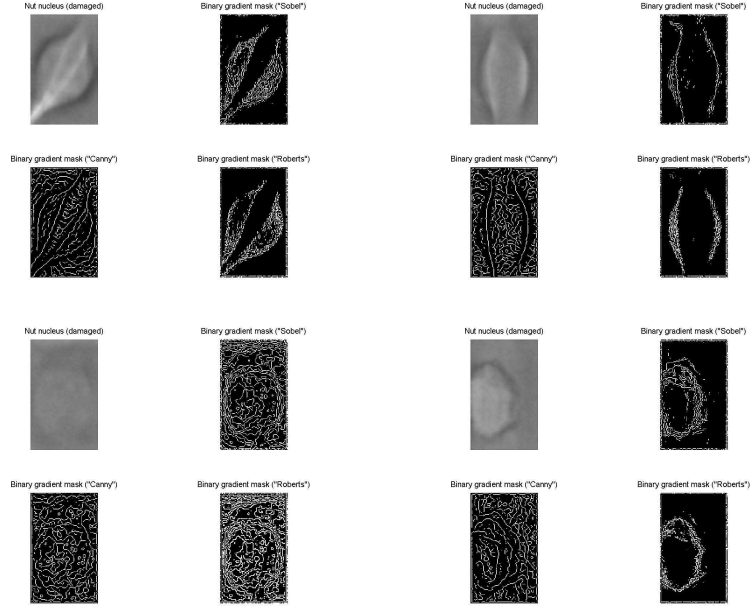


Figure 5: Image processing of the hazelnut nuclei belonging to the set D , for $\epsilon = 100$ and $\rho = 1.5$, by means of edge-detection algorithms, respectively: Sobel's algorithm (top right figure) , Canny's algorithm (bottom left figure) and Roberts' algorithm (bottom right figure).

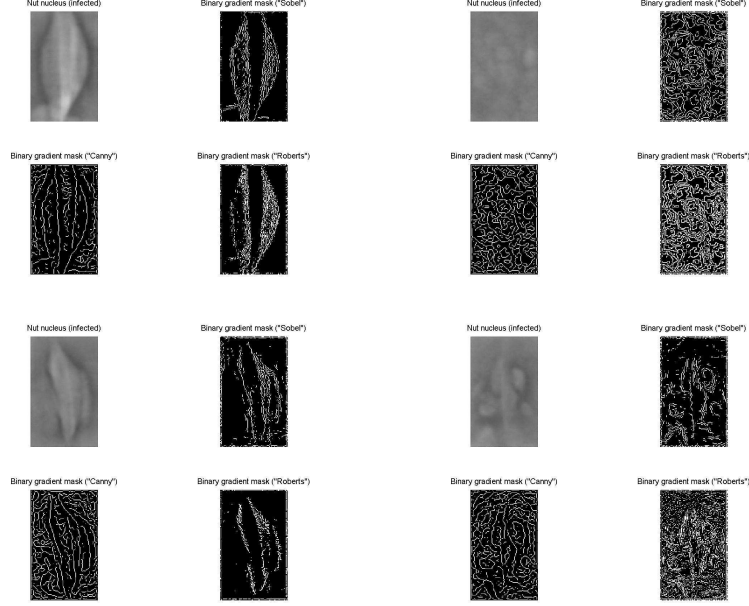


Figure 6: Image processing of the hazelnut nuclei belonging to the set I , for $\epsilon = 100$ and $\rho = 1.5$, by means of edge-detection algorithms, respectively: Sobel's algorithm (top right figure) , Canny's algorithm (bottom left figure) and Roberts' algorithm (bottom right figure).

In Fig. 7, which is worth comparing with Fig. 2, we plotted the mean histograms relative to the cropped images, with $\epsilon = 80$ and $\rho = 2.5$. The question arises, then, as to whether the separation between the two scales $\langle d \rangle^{(A)}$ and $\Delta^{(A,B)}$ is amenable to be enhanced by tuning the two parameters ϵ and ρ .

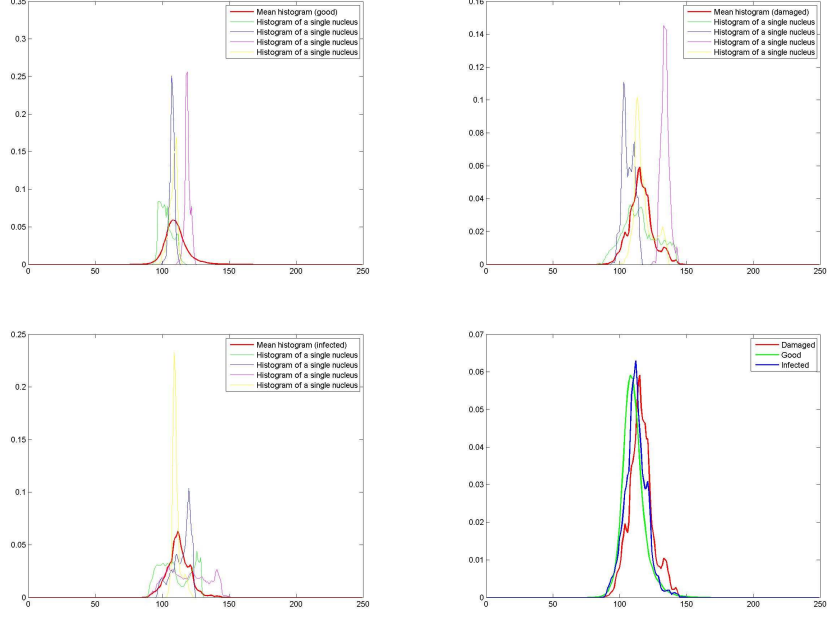


Figure 7: Image histograms of the hazelnut nuclei belonging to the sets G (top left), D hazelnuts (top right), I hazelnuts (bottom left). For each of the three sets, the figures display the histograms of single hazelnuts as well as the mean histogram in the corresponding set (mean histogram). On the bottom right corner, the mean histograms of the three sets are compared. All the histograms were obtained by setting $\epsilon = 80$ and $\rho = 2.5$.

We thus studied the behaviour of the mean histograms, shown in Fig. 7, as well as of the typical fluctuations occurring in each set, as a function of ϵ and ρ : in our simulations, ϵ spans a broad range of values, whereas we let ρ attain the values 1.5 and 2.5, cf. Fig. 8.

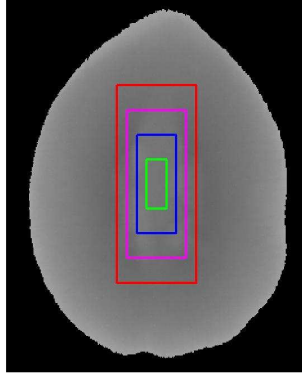


Figure 8: Different values of the scale of resolution: $\epsilon = 80$ (red rectangle), $\epsilon = 60$ (magenta rectangle), $\epsilon = 40$ (blue rectangle), $\epsilon = 20$ (green rectangle). All the colored rectangles shown in the picture are obtained by setting $\rho = 2.5$.

In Fig. 9 and 10, the mean histograms of the sets G , D and I are shown for different values of ϵ , and for two different values of ρ .

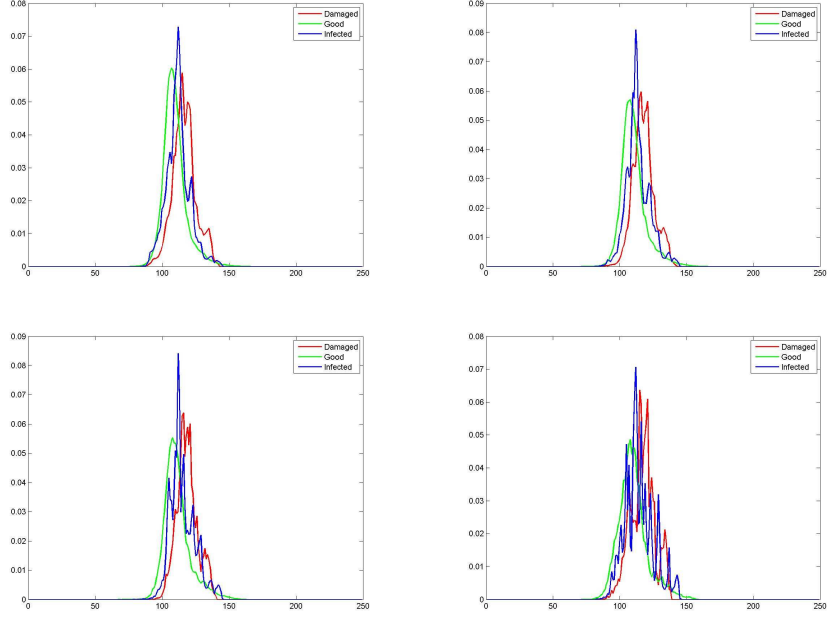


Figure 9: Mean histograms of the hazelnut nuclei at different scales of resolution: $\epsilon = 80$ (top left), $\epsilon = 60$ (top right), $\epsilon = 40$ (bottom left) and $\epsilon = 20$ (bottom right), with $\rho = 1.5$.

We focused, in particular, on the investigation of the dependence of the scales $\langle d \rangle^{(A)}(\epsilon; \rho)$ and $\Delta^{(A,B)}(\epsilon; \rho)$ on the resolution ϵ .

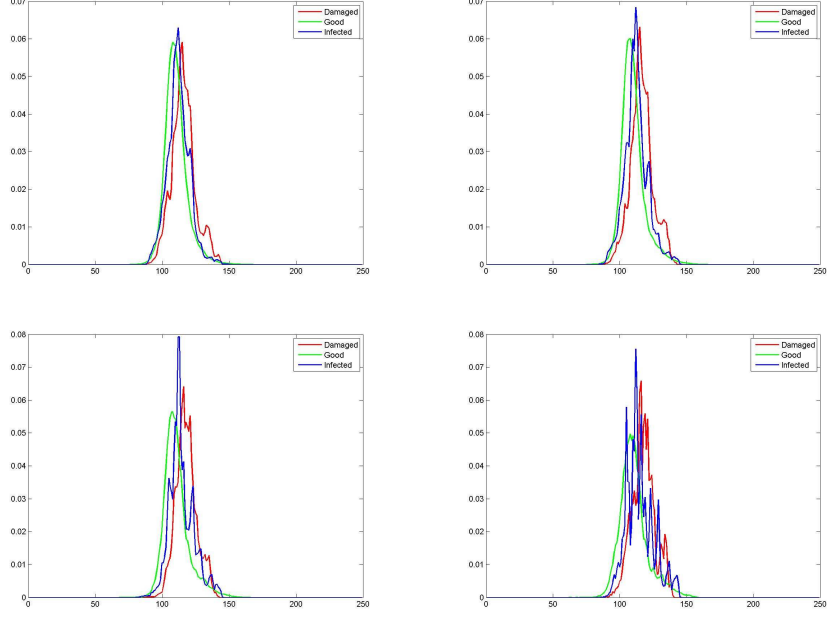


Figure 10: Mean histograms of the nuclei of the hazelnuts at different scales of description: $\epsilon = 80$ (top left), $\epsilon = 60$ (top right), $\epsilon = 40$ (bottom left) and $\epsilon = 20$ (bottom right), with $\rho = 2.5$.

Figures 11 and 12 illustrate the behaviour of $\langle d \rangle^{(A)}$ and $\Delta^{(A,B)}$ vs. ϵ for $\rho = 1.5$, whereas Figs. 13 and 14 show the analogous behaviour of $\langle d \rangle^{(A)}$ and $\Delta^{(A,B)}$ vs. ϵ for $\rho = 2.5$

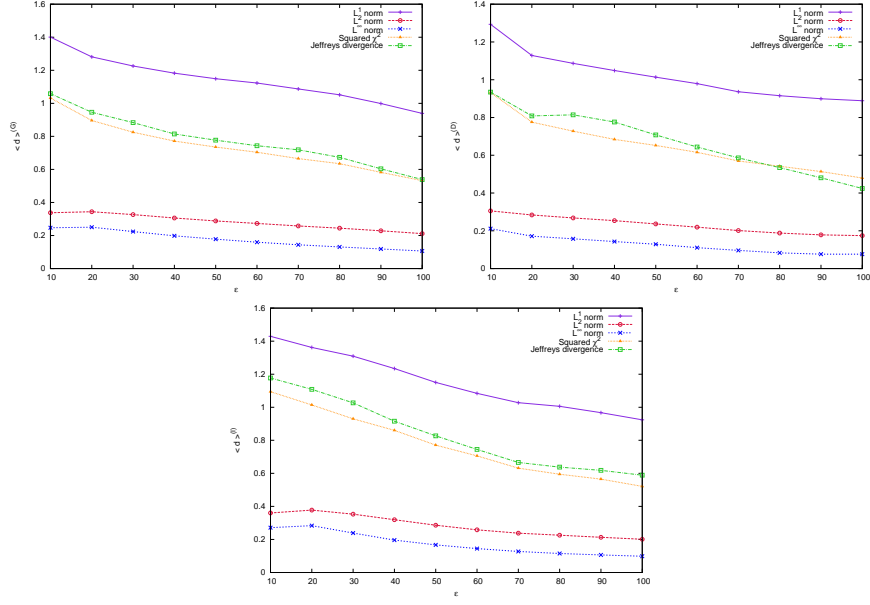


Figure 11: Behaviour of the distances $\langle d \rangle^{(A)}$ vs. ϵ , with $\rho = 1.5$.

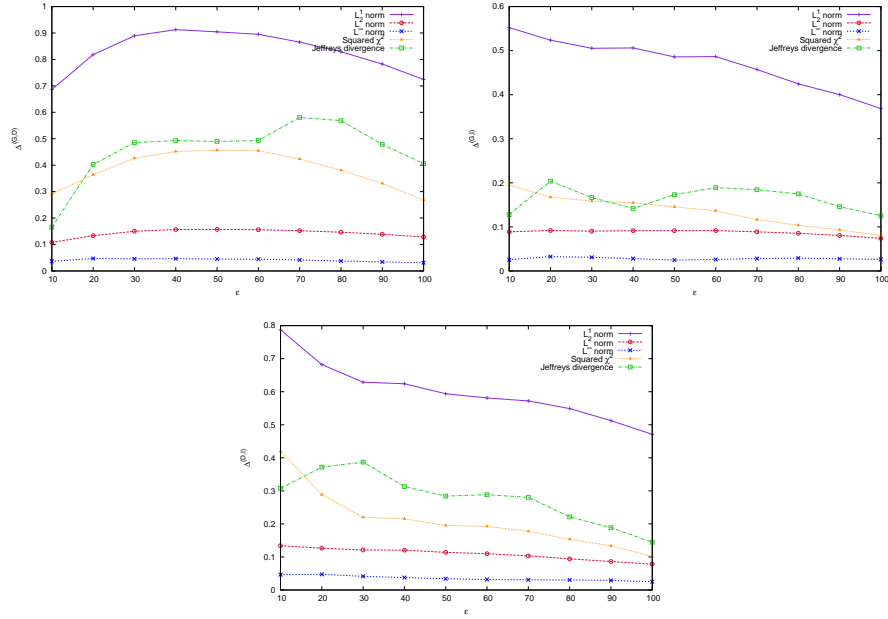


Figure 12: Behaviour of the distances $\Delta^{(A,B)}$ vs. ϵ , with $\rho = 1.5$.

The two plots 12 and 14 reveal that reducing ϵ leads, on the one hand, to a remarkable increase of $\Delta^{(A,B)}$, which attains an order of magnitude of about $\Delta^{(A,B)} \simeq 10^{-1}$. On the other hand, this effect is counterbalanced by the simultaneous increase of the scale $\langle d \rangle^{(A)}$, evidenced in Figs. 11 and 13, which turns out to be, for both the considered values of ρ , of the same order of magnitude of $\Delta^{(A,B)}$. This is more clearly visible in Fig. 15, which illustrates the behaviour of the ratio of $\Delta^{(A,B)}$ to $\langle d \rangle^{(A)}$ and to $\langle d \rangle^{(B)}$, for different values of ϵ , obtained by setting $A = G$ and $B = D$.

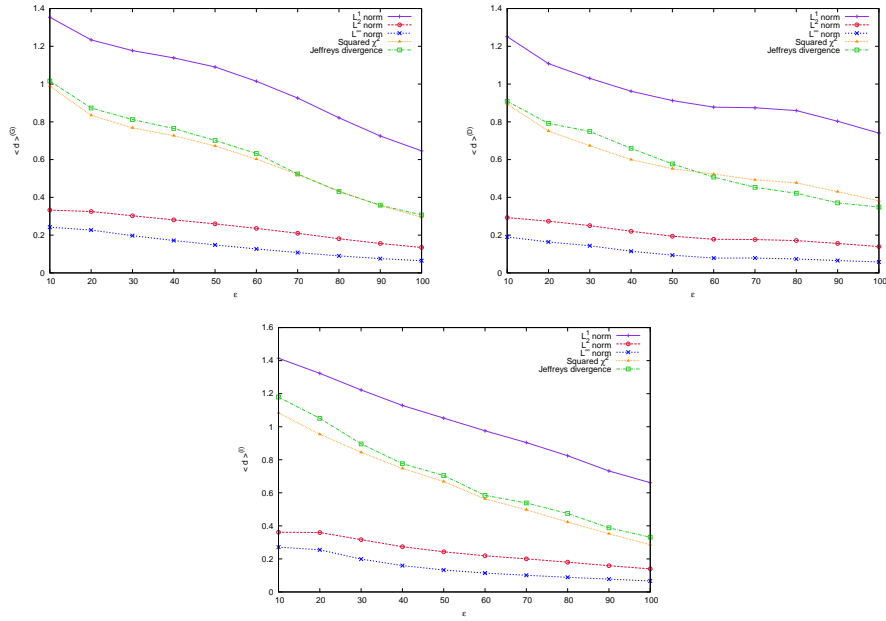


Figure 13: Behaviour of the distances $\langle d \rangle^{(A)}$ vs. ϵ , with $\rho = 2.5$.

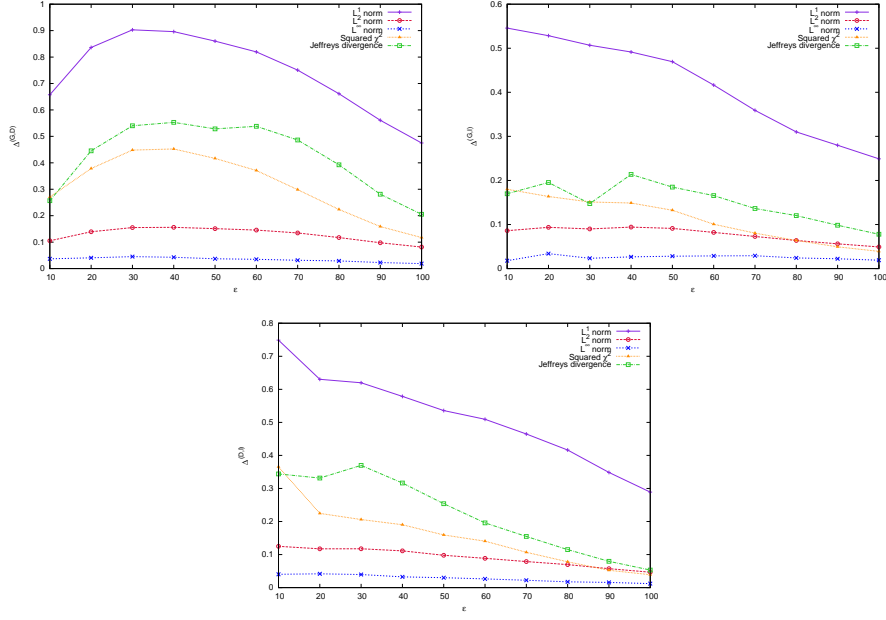


Figure 14: Behaviour of the distances $\Delta^{(A,B)}$ vs. ϵ , with $\rho = 2.5$.

The plots in Fig. 15 confirm that the two scales $\Delta^{(A,B)}$ and $\langle d \rangle^{(A)}$ remain of the same order, also when reducing ϵ . On the contrary, an efficient pattern recognition, based on the analysis of the image histograms, can be obtained if the ratio $\Delta^{(A,B)}/\langle d \rangle^{(A)} \gg 1$, i.e. when the mean statistical distance between different sets overwhelms the typical size of fluctuations characteristic of each set. Thus, the study of the behaviour of the two latter scales allows one to predict a poor performance of a machine learning algorithm aiming at classifying the hazelnuts on the basis of the image histograms. Nevertheless, an interesting aspect can be evinced from an inspection of Fig. 15: despite the similarity of the magnitudes of the two statistical scales, the plot of their ratio vs. ϵ yields a non monotonic function.

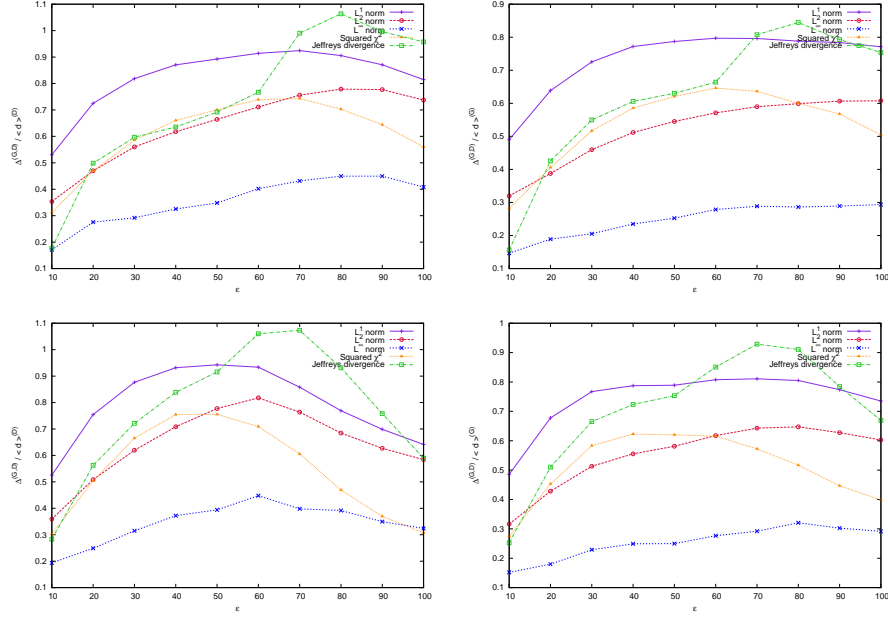


Figure 15: Behaviour of the ratio $\Delta^{(G,D)}/\langle d \rangle^{(G)}$ (left column) and $\Delta^{(G,D)}/\langle d \rangle^{(D)}$ (right column) vs. ϵ , for $\rho = 1.5$ (upper row) and $\rho = 2.5$ (lower row).

To better evidence this point, we plotted, in Fig. 16, the ratio of the scale $\Delta^{(A,B)}$ to the geometric mean $\sqrt{\langle d \rangle^{(A)} \langle d \rangle^{(B)}}$, where we set $A = G, B = D$ (left panel) and $A = G, B = I$ (right panel).

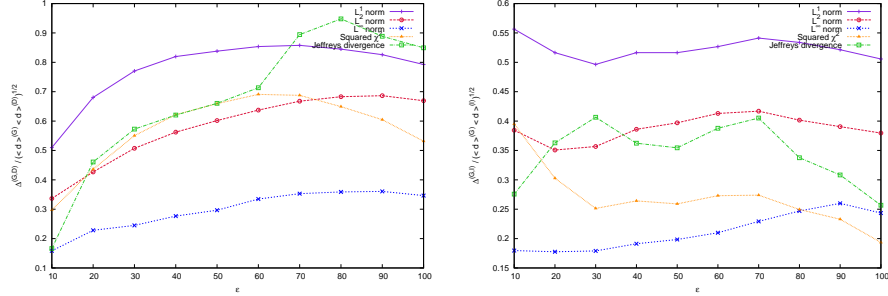


Figure 16: *Left panel:* Behaviour of the ratio $\Delta^{(G,D)}/\sqrt{\langle d \rangle^{(G)} \langle d \rangle^{(D)}}$ vs. ϵ , for $\rho = 1.5$. *Right panel:* Behaviour of the ratio $\Delta^{(G,I)}/\sqrt{\langle d \rangle^{(G)} \langle d \rangle^{(I)}}$ vs. ϵ , for $\rho = 1.5$.

In Fig. 17, instead, for reasons to be further clarified in Sec. 3, we show the results, analogous to those portrayed in Fig. 16, obtained by merging the two sets D and I into one single set, labeled as nG (“not good” hazelnuts).

The plot in Fig. 17 shows that, for $\rho = 1.5$, the value $\epsilon^* = 70$ maximizes the ratio of the aforementioned statistical scales wrt almost all the various notions of “statistical distance” we considered. In Sec. 3 we will show that such optimal value ϵ^* , here obtained by only relying on information theoretic methods, can be also recovered by using Support Vector Machines numerical algorithms, by averaging their performance over a set of training samples.

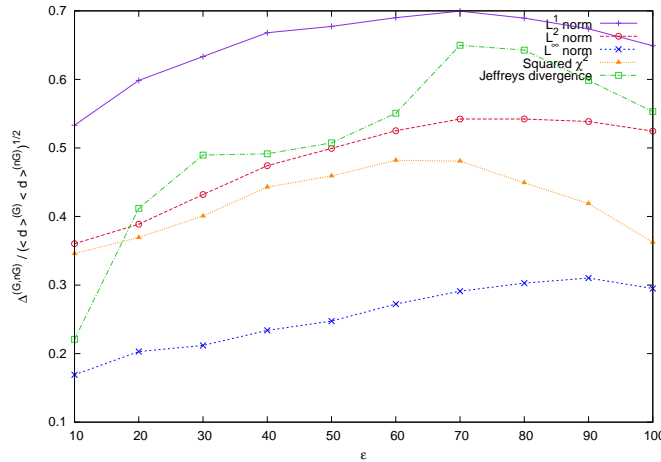


Figure 17: *Left panel:* Behaviour of the ratio $\Delta^{(G, nG)} / \sqrt{\langle d \rangle^{(G)} \langle d \rangle^{(nG)}}$ vs. ϵ , for $\rho = 1.5$. The plot evidences the onset of an optimal scale ϵ^* at which the ratio of the statistical scales is maximized.

3. Support Vector Machines

In this Section, we discuss the results obtained by elaborating our data through a supervised learning method known as Support Vector Machines (SVM) [15, 28, 29, 30, 31]. The SVM constitute a machine learning algorithm which seeks a separation of a set of data into two classes, by determining the *best separating hyperplane* (BSH) (also referred to, in the literature, as the “maximal margin hyperplane” [16]), cf. Fig. 18. It is worth recapitulating the basic notions underpinning the numerical algorithm we used.

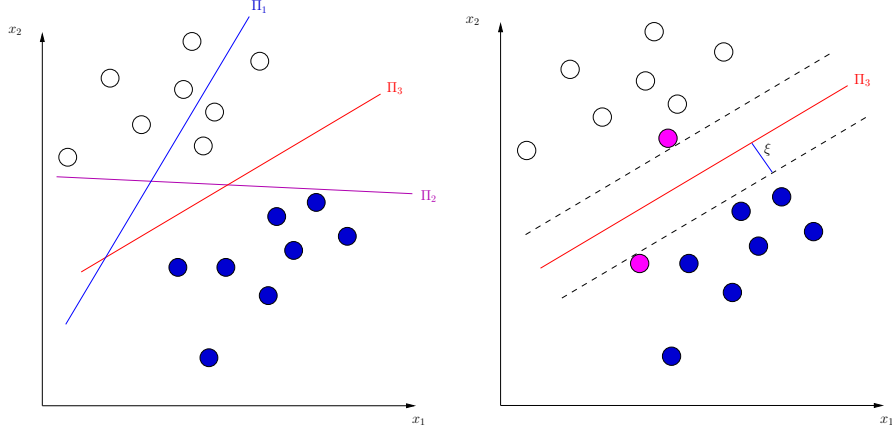


Figure 18: *Left panel:* Example of a linear discriminant analysis based on the SVM algorithm. Shown are three different hyperplanes: Π_1 , which does not separate the two classes, Π_2 which separates the classes but only with a small margin, and Π_3 , which corresponds to the best separating hyperplane. *Right panel:* Illustration of the best separating hyperplane (red straight line), the canonical hyperplanes (black dashed lines), the support vectors (magenta circles) and the margin of separation ξ .

Let $\{\mathbf{x}\}$ denote the set of data (input pattern) to be classified, with $\mathbf{x} \in E \subseteq \mathbb{R}^N$, and consider a given training set $\mathcal{T} = \{\mathbf{x}_k, d_k\}_{k=1}^{N_T}$, where N_T denotes the dimensionality of \mathcal{T} . Let, then, $d_k = \{+1, -1\}$ denote the *desired response* parameter corresponding to \mathbf{x}_k , whose value depends on which of the two classes \mathbf{x}_k belongs to. The equation of a hyperplane Π in \mathbb{R}^N reads:

$$\mathbf{w}^T \cdot \mathbf{x} + b = 0 \quad ,$$

with \mathbf{w} and b denoting, respectively, a N -dimensional adjustable weight vector and a bias. The BHS is the hyperplane characterized by the pair (\mathbf{w}_o, b_o) which, for linearly separable patterns, fulfills the following conditions [15]:

$$\begin{aligned} \mathbf{w}_o^T \cdot \mathbf{x}_k + b_o &\geq 1 \quad \text{for } d_k = +1 \quad , \\ \mathbf{w}_o^T \cdot \mathbf{x}_k + b_o &\leq -1 \quad \text{for } d_k = -1 \quad . \end{aligned} \quad (6)$$

The data points, portrayed in magenta color in the right panel of Fig. 18, for which Eqs. (6) are satisfied with the equality sign, are called *support vectors*, and lie on the so-called *canonical hyperplanes* [16], represented by the black dashed lines in the right panel of Fig. 18. Figure 18 also illustrates the so-called *margin of separation*, defined as the distance $\xi = 1/\|\mathbf{w}_o\|$

between the support vectors and the BSH. The BSH, which maximizes ξ under the constraints (6), can be found by determining the saddlepoint of the Lagrangian function $d\mathcal{L}(\mathbf{w}, b, \lambda_1, \dots, \lambda_{N_T}) = 0$, given by:

$$\mathcal{L}(\mathbf{w}, b, \lambda_1, \dots, \lambda_{N_T}) = \frac{1}{2} \mathbf{w}^T \cdot \mathbf{w} - \sum_{k=1}^{N_T} \lambda_k [d_k(\mathbf{w} \cdot \mathbf{x}_k + b) - 1] \quad . \quad (7)$$

The solution of such variational problem is easily found in the form [15]:

$$\mathbf{w}_o = \sum_{k=1}^{N_T} \lambda_k d_k \mathbf{x}_k \quad (8)$$

where the Lagrange multipliers λ_k satisfy the conditions:

$$\begin{aligned} \sum_{k=1}^{N_T} \lambda_k d_k &= 0 \quad , \\ \lambda_k [d_k(\mathbf{w} \cdot \mathbf{x}_k + b) - 1] &= 0 \quad \text{for } k = 1, \dots, N_T \quad , \end{aligned}$$

(the latter being known as the “Karush-Kuhn-Tucker complementarity condition” [16]) whereas b_o can be determined, once \mathbf{w}_o is known, using Eqs. (6). When the two classes are not linearly separable, a possible strategy consists in introducing a suitable (nonlinear) function $\Phi : E \rightarrow F$, which makes it possible to map the original pattern inputs into a *feature space* $F \subseteq \mathbb{R}^M$, in which a linear separation can be performed, cf. Fig. 19 [16].

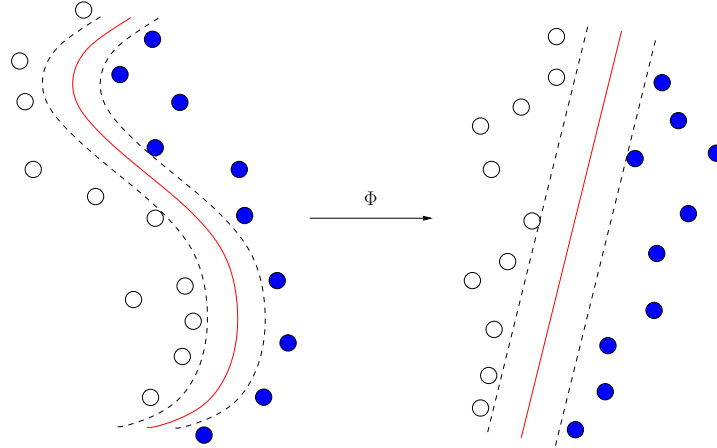


Figure 19: Patterns which are not linearly separable can be mapped, via a function Φ , into a *feature space* where a linear separation of the classes can be achieved

Thus, by denoting as $\Phi(\mathbf{x}) = \{\Phi_j(\mathbf{x})\}_{j=1}^M$ a set of nonlinear transformations from the original input space to the feature space, the corresponding variational problem leads now, in place of Eq. (8), to the expression:

$$\mathbf{w}_o = \sum_{k=1}^{N_T} \lambda_k d_k \Phi(\mathbf{x}_k) \quad . \quad (9)$$

In our implementation of the SVM algorithm, we regarded the set G as one of the two classes, whereas the other class, formerly introduced in Sec. 2 and denoted by nG , was thought of as given by the union $nG = D \cup I$. We thus relied on the analysis of the histograms of the hazelnut nuclei, detailed in Sec. 2. Therefore, we introduced two variables to identify each hazelnut: we set $\mathbf{x} = (x_{mean}, x_{max})$, where, for each histogram relative to an hazelnut nucleus, x_{mean} and x_{max} denote, respectively, the *average* shade of gray and the shade of gray equipped with the highest probability. Therefore, in the space spanned by the coordinates x_{mean} and x_{max} , and parameterized by the values of ϵ and ρ , each hazelnut is represented by a single dot. The resulting distribution of dots, for different values of ϵ and ρ , is illustrated in Figs. 20 and 21, which evidence a clustering of points, for both the considered values of ρ , around the bisectrix of the plane. This is readily explained by considering that, when reducing ϵ , the histograms attain a more and more symmetric shape.

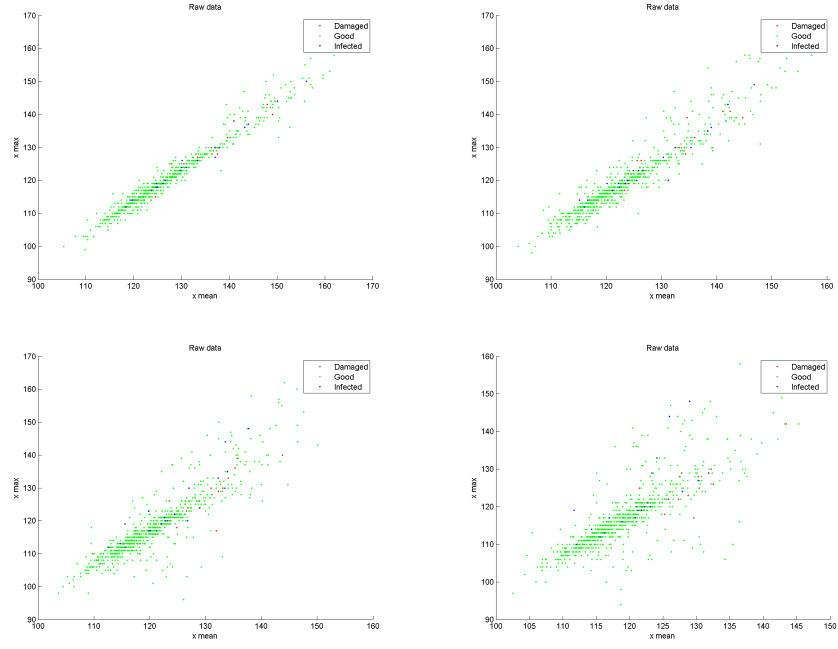


Figure 20: Classification of the data in the 2D space spanned by the values of the observables x_{mean} (horizontal axis) and x_{max} (vertical axis), for $\epsilon = 20$ (top left), $\epsilon = 40$ (top right), $\epsilon = 60$ (bottom left), and $\epsilon = 80$ (bottom right), with $\rho = 1.5$.

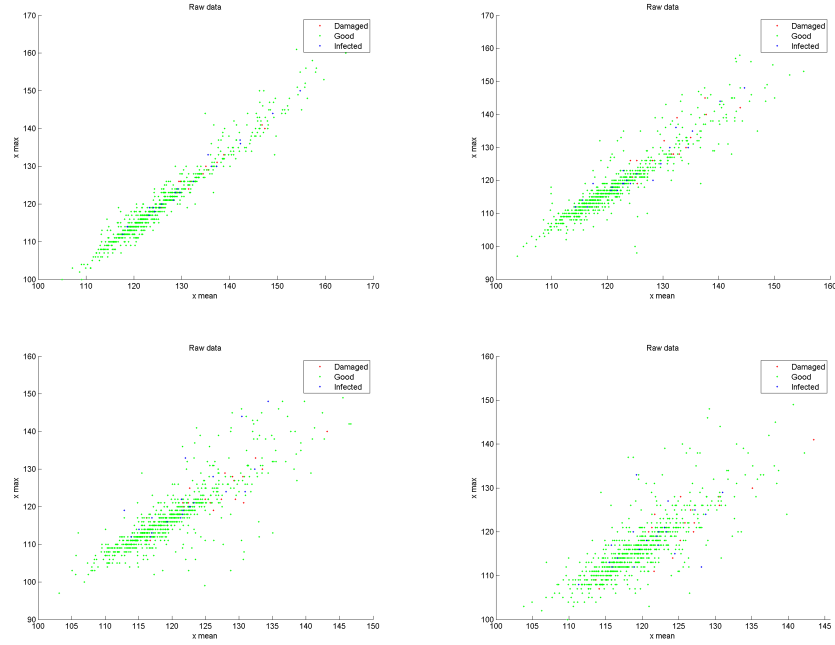


Figure 21: Classification of the data in the 2D space spanned by the values of the observables x_{mean} (horizontal axis) and x_{max} (vertical axis), for $\epsilon = 20$ (top left), $\epsilon = 40$ (top right), $\epsilon = 60$ (bottom left), and $\epsilon = 80$ (bottom right), with $\rho = 2.5$.

Furthermore, an inspection of Figs. 20 and 21 reveals that the dots corresponding to the sets D and I are nested within the ensemble of points belonging to the set G : the classes G and nG are not amenable to be disentangled by a linear SVM regression, as also confirmed by the plots in Figs. 22 and 23. In each of the two latter figures, the left plot shows the elements of the adopted (randomly selected) training set: green and red symbols identify the elements of the two classes G and nG , while the black circles indicate the support vectors. The black line indicates the boundary (best separating hyperplane) detected by the SVM, which sensibly depends on the chosen training set. The right plot, instead, displays all the available data (red and blue crosses represent, respectively, the elements of the classes G and nG), complemented by the SVM test set output (red and blue circles). The proper match between the colours of the circles and the crosses would indicate a successfully accomplished separation between the two classes, which, though, is not obtained with our data. Furthermore, no remarkable improvement is obtained by attempting a classification of the data by means of a nonlinear SVM algorithm, based on radial basis func-

tions [16], as shown in Figs. 24 and 25. The results of this Section, confirm, therefore, the predictions of the statistical analysis outlined in Sec. 2: the presence of a not linearly separable entanglement between points belonging to different classes can be thus traced back to the lack of a suitable statistical scales separation.

There is another relevant aspect, concerned with the implementation of the SVM algorithm, to be pointed out.

We remark, in fact, that each of the plots shown in Figs. 22 and 23 pertains to a specific training set of data \mathcal{T} .

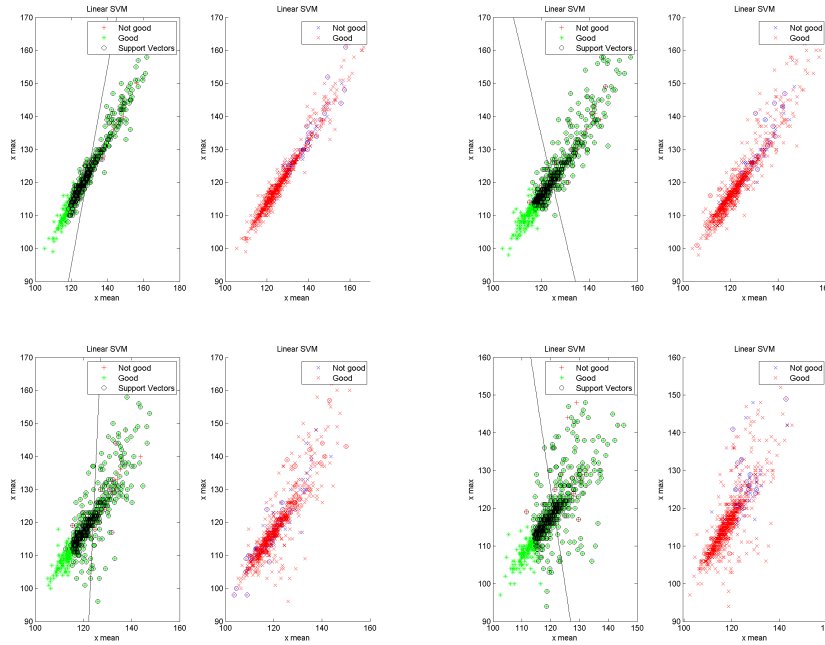


Figure 22: Classification of the data through a linear SVM algorithm. Shown is the 2D space spanned by the values of the observables x_{mean} (horizontal axis) and x_{max} (vertical axis), for $\epsilon = 20$ (top left), $\epsilon = 40$ (top right), $\epsilon = 60$ (bottom left), and $\epsilon = 80$ (bottom right), with $\rho = 1.5$. In each left subfigure, shown are the training set of data (green and red crosses, denoting, respectively, the elements of the classes G and nG), the support vectors (black circles) and the best separating hyperplane (black line). According to the SVM classification, the elements of the class nG are expected to lie on the right of the boundary line. The right sub-figures, instead, display the 2D representation of all the available data (red and blue crosses, denoting, respectively, the elements of G and those of nG) and the SVM output (red and blue circles).

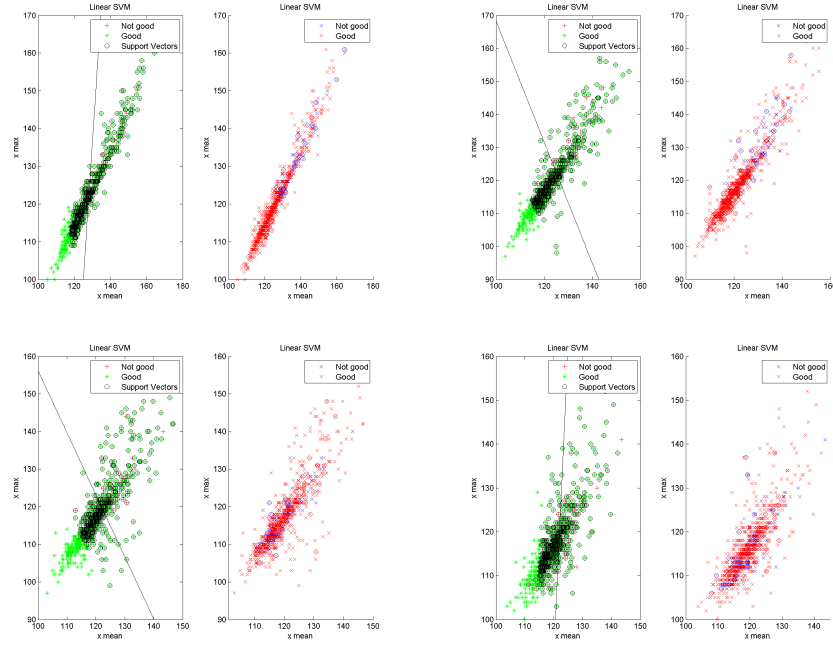


Figure 23: Classification of the data with a linear SVM algorithm, as in Fig. 22, but with $\rho = 2.5$.

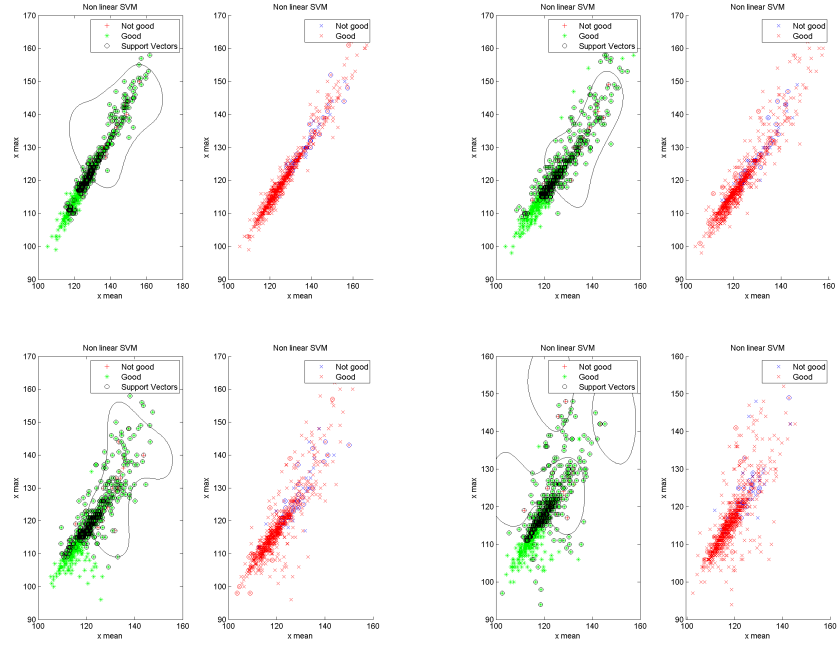


Figure 24: Classification of the data through a nonlinear SVM algorithm (based on radial basis functions) for $\rho = 1.5$ (cf. the caption of Fig. 22).

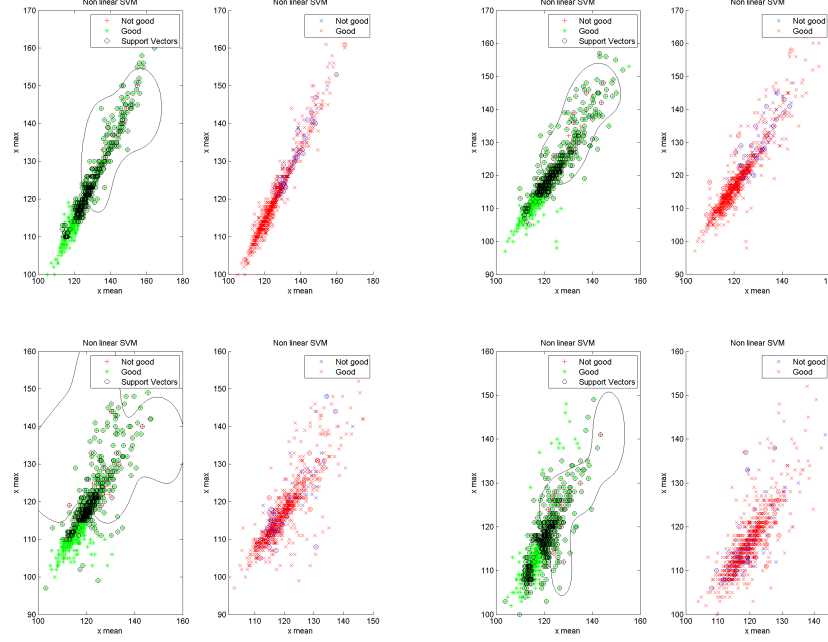


Figure 25: Classification of the data with a nonlinear SVM algorithm (based on radial basis functions) for $\rho = 2.5$ (cf. the caption of Fig. 22).

We can introduce, then, the quantity $\Psi_\ell(\mathcal{T}_\ell; \epsilon, \rho)$, relative to the specific training set \mathcal{T}_ℓ , and defined as the ratio of the number of hazelnuts, belonging to the class G and mistakenly classified as belonging to the class nG , to the total number of hazelnuts in the database, given by $N_G + N_{nG}$. The function Ψ_ℓ is an indicator of the performance of the SVM algorithm, and sensibly depends on the structure of the training sample considered in the simulation. Thus, while the behaviour of Ψ_ℓ , pertaining to single training samples, yields no indication about the onset of an optimal scale ϵ^* , the average $\langle \Psi \rangle$, given by

$$\langle \Psi \rangle(\epsilon, \rho) = \frac{1}{N_c} \sum_{\ell=1}^{N_c} \Psi_\ell(\mathcal{T}_\ell; \epsilon, \rho) \quad ,$$

and computed over a sufficiently large number N_c of training samples, attains a minimum precisely at $\epsilon^* = 70$, cf. Fig. 26. The latter value of ϵ corresponds, in fact, to the scale of resolution maximizing the two statistical scales introduced in Sec. 2, cf. Fig. 17. The plot in Fig. 26 confirms, hence, that the onset of an optimum scale ϵ^* can be numerically evinced also by means of SVM algorithms, provided that performance of the SVM

is *averaged* over a sufficiently large number of training samples.

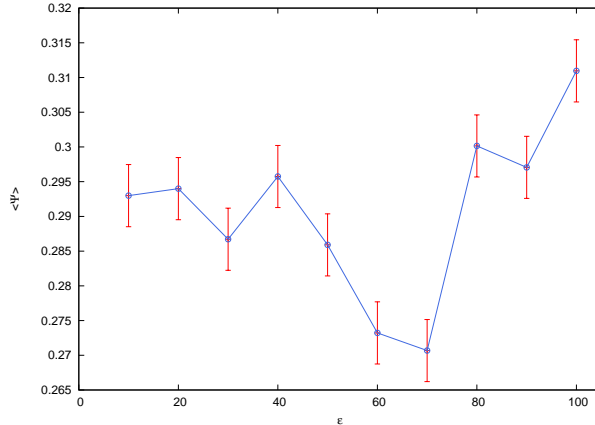


Figure 26: Behavior of $\langle \Psi \rangle$, averaged over $N_c = 500$ training samples, vs. ϵ , with $\rho = 1.5$, with error bars (in red).

4. Conclusions

In this work we performed a statistical analysis on the histograms of a set of hazelnut images, with the aim of obtaining a preliminary estimate of the performance of a machine learning algorithm based on statistical variables. We shed light, in Sec. 2, on the relevance of two statistical scales, which need to be widely separated to accomplish a successful pattern recognition. The intrinsic lack of such scale separation in our data was also evidenced by the numerical results reported in Sec. 3, revealing that no exhaustive classification can be achieved through SVM algorithms. Moreover, the analysis outlined in Sec. 2 also unveiled the onset of an optimal resolution ϵ^* , which is expected to optimize the pattern recognition. This observation was also corroborated by the results discussed in Sec. 3, where the same value ϵ^* , maximizing the performance of the SVM algorithm, is recovered by averaging over a sufficiently large number of training samples. Our results, thus, strengthen the overall perspective that a preliminary estimate of the intrinsic statistical scales of the data constitute a decisive step in the field of pattern recognition and, moreover, pave the way for the further implementation of statistical mechanical techniques aimed at the development of a generation of more refined neural networks algorithms.

Acknowledgments

We would like to thank Ferrero and Soremartec for their long-standing support of our research activity. We also thank Dr. A. Boscolo and Dr. L. Placentino, for providing us with the set of x-ray images used in this work. This study was funded by ITACA, a project financed by the European Union, the Italian Ministry of Economy and Finance and the Piedmont Region.

References

- [1] M. Egmont-Petersen, D. de Ridder, H. Handels, Image processing with neural networks - a review, *Pattern Recognition* **35**, 2279 (2002).
- [2] T. Wang, X. Zhuang, X. Xing, Robust segmentation of noisy images using a neural network model, *Im. Vision Comput.* **10**, 233 (1992).
- [3] Y. Hayashi, R. Setiono, Combining neural network predictions for medical diagnosis, *Comp. in Biol. and Med.* **32**, 237 (2002).
- [4] J. Jiang, P. Trundle, J. Ren, Medical Image Analysis with artificial Neural Network, *Comput. Med. Imaging Graph* **34**, 617 (2010).
- [5] Q. K. Al-Shayea, Artificial Neural Networks in Medical Diagnosis, *Int. J. Comp. Sc. Is.* **8**, 150 (2011).
- [6] C. H. Q. Ding, I. Dubchak, Multi-class protein fold recognition using support vector machines and neural networks, *Bioinformatics* **17**, 349 (2001).
- [7] L. Qian, E. Winfree, J. Bruck, Neural network computation with DNA strand displacement cascades, *Nature* **475**, 368 (2011).
- [8] A. Condon, Molecular programming: DNA and the brain, *Nature* **475**, 304 (2011).
- [9] H. M. Cartwright, Artificial Neural Networks in Biology and Chemistry - The Evolution of a New Analytical Tool, *Art. Neur. Netw., Methods in Molecular Biology* **458**, 1 (2009).
- [10] D. J. C. MacKay, Information Theory, Interference, and Learning Algorithms (Cambridge University Press, 2003).

- [11] K. Anand, G. Bianconi, Entropy measures for networks: Toward an information theory of complex topologies, *Phys. Rev. E* **80** 045102 (R) (2009).
- [12] G. Tkacik, O. Marre, T. Mora, D. Amodei, M. J. Berry II, W. Bialek, The simplest maximum entropy model for collective behavior in a neural network, *J. Stat. Mech* (2013), P03011.
- [13] K.-J. Kim, S.-B. Cho, Diverse Evolutionary Neural Networks Based on Information Theory, *Neural Information Processing, Lecture Notes in Computer Science* **4985**, 1007 (2008).
- [14] N. X. Vihn, J. Bailey, Information Theoretic Measures for Clustering Comparison: Variants, Properties, Normalization and Correction for Chance, *J. Mach. Learn. Res.* **11**, 2837 (2010).
- [15] S. Haykin, Neural Networks. A Comprehensive Foundation (Prentice-Hall, 1999).
- [16] A. Webb, Statistical Pattern Recognition (John Wiley and Sons, 2002).
- [17] S.-H. Cha, S. N. Srihari, On measuring the distance between histograms, *Pattern Recognition* **35**, 1355 (2002).
- [18] S.-H. Cha, Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions, *International Journal of Mathematical Models and Methods in Applied Sciences* **1**, 300 (2007).
- [19] R. L. Plackett, Karl Pearson and the Chi-Squared Test, *International statistical Review* **51**, 59 (1983).
- [20] H. Jeffreys, An Invariant Form for the Prior Probability in Estimation Problems, *Proc. Roy. Soc. Lon.*, Ser. A, **186**, 453 (1946).
- [21] C. Beck, Generalized information and entropy measures in physics, *Contemporary Physics* **50**, 495 (2009).
- [22] S. Kullback, R.A. Leibler, On information and sufficiency, *Ann. Math. Statist.* **22**, 79 (1951).
- [23] S. Kullback, Information Theory and Statistics (Dover Press, 1959).
- [24] E. T. Jaynes, Information theory and statistical mechanics, *Physical Review* **106**, 620 (1957).

- [25] T. Morimoto, Markov processes and the H-theorem, *J. Phys. Soc. Jap.* **18** 328 (1963).
- [26] S. M. Ali, S. D. Silvey, A general class of coefficients of divergence of one distribution from another, *Journal of the Royal Statistical Society, Ser. B*, **28**, 131 (1966).
- [27] D. Xu, D. Erdogmus, Renyi's Entropy, Divergence and Their Nonparametric Estimators, in *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives* (Springer 2010).
- [28] B. E. Boser , I. M. Guyon and V. N. Vapnik, A Training Algorithm for Optimal Margin Classifiers, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, 144 (1992).
- [29] C. Cortes and V. N. Vapnik, Support vector networks, *Machine Learning* **20**, 273 (1995).
- [30] V. N. Vapnik, The Nature of Statistical Learning Theory (Springer-Verlag, New York, 1995).
- [31] V. N. Vapnik, Statistical Learning Theory (Wiley, New York, 1998).